DOCUMENT RESUME

ED 049 292 TM 000 479

AUTHOR Greene, John F.
TITLE Computer Simulation of Human Behavior: Assessment of Creativity.
PUB DATE Feb 71
NOTE 17p.; From the Symposium on "Multiple Regression Prediction Models in the Behavioral Sciences"; presented at the Annual Meeting of the AERA, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Behavioral Science Research, Computer Programs, Correlation, Creative Ability, *Creativity Research, Creativity Tests, *Models, Multiple Regression Analysis, Predictive Measurement, Predictor Variables, *Rating Scales, Reliability, Research Methodology, Scoring, *Simulation, *Statistical Analysis, Validity
IDENTIFIERS *Torrance Tests of Creative Thinking, TTCT

ABSTRACT

The major purpose of this study is to further the development of procedures which minimize current limitations of creativity instruments, thus yielding a reliable and functional means for assessing creativity. Computerized content analysis and multiple regression are employed to simulate the creativity ratings of trained judges. The computerized scoring procedure is evaluated in a cross-validation sample. The methodological problems of establishing a reliable criterion and generating parsimonious forced prediction models through predictor stability analysis are emphasized, and possible solutions are explored. Support for the proposed solutions and empirical validation are incorporated in the study. The forced model results are regarded as tentative and should be tested on a new sample. Bibliography and statistical data are included. (Author/AE)

ED049292

COMPUTER SIMULATION OF HUMAN BEHAVIOR: ASSESSMENT OF CREATIVITY

John F. Greene
The University of Bridgeport

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

TM 000 429

Presented at the symposium on
Multiple Regression Prediction Models in the Behavioral Sciences

AERA Annual Meeting, 1971
New York, New York

Computer Simulation of Human Behavior:

Assessment of Creativity

John F. Greene
The University of Bridgeport

This report is divided into three sections. In Section I a review of the basic research study is presented. The study represents the third stage of ongoing research in the field of scoring creativity tests by computer. In stage one Dieter Paulus and Joseph Renzulli generated the idea of conducting such research and demonstrated feasibility. Computerized scoring procedures for three creativity tasks were developed by Francis Archambault during the second stage. The last two sections of this report consider the methodological problems of establishing reliable criteria and generating parsimonious prediction models in the behavioral sciences as related to this study.

I. Review of the Original Research Study

The major purpose of this study was to further the development of procedures which minimize the current limitations of creativity instruments, thus yielding a more reliable and functional means for assessing creativity. computerized content analysis was employed to simulate the creativity ratings that trained human judges make in the process of scoring the free, open-ended responses to Torrance Tests of Creative Thinking (TTCT) (Torrance, 1966a). The Verbal, Form A version of the TTCT served as the basic source upon which reliable and functional scoring strategies were developed, but these strategies are not necessarily limited to this test battery. Form A of the TTCT consists of seven activities. Only the last four, however, were considered in this study. Activities four through seven are Product Improvement (toy elephant), Unusual Uses (cardboard boxes), Unusual Questions (cardboard boxes), and Just Suppose (if clouds had strings, what would happen?) respectively.

Each activity is scored for three dimensions of creativity: fluency, flexibility, and originality. A flexibility score, however, is not determined for the sixth activity, Unusual Questions.

The scoring procedures remain constant throughout the first five activities of the TTCT. Different techniques, however, are employed in activities six and seven. The computerized scoring strategies for activities four and five included elaborate dictionaries. These dictionaries were constructed using the categories provided by Torrance as the basic structure. Activities six and seven present the unique challenges of determining the complexity of the answer given only the question and detecting shifts in attitude or focus between responses respectively. The computerized scoring procedures developed for these two tasks are beyond the scope of this paper. Acturial variables were employed to supplement prediction in all four activities.

From a sample of 375 students used in a study by Treffinger and Ripple (1968), 153 subjects were randomly selected. Four judges rated the responses of each subject. Analysis of variance procedures were used to provide a reliability estimate of the pooled ratings of the judges (Winer, 1962, pp. 124-132). A step-wise multiple regression technique was employed to maximize the prediction of each subject's scores for each activity. The predictors included the acturial and dictionary parameters generated earlier by the SCRTXT computer program (Fisher, 1968). Besides the full model, restricted and forced regression models were generated. The entire computerized scoring procedure was then evaluated in a cross-validation sample.

The adjusted pooled reliability estimates based on all four judges for fluency and flexibility were most satisfactory, ranging from .88 to .99 with 6 of 7 above .94. The originality reliabilities, although satisfactory, were somewhat lower. Their range was bounded by .66 and .86.

The results of the multiple correlation analyses are summarized in Table I. The full model coefficients for fluency range from .92 to .99.

The range for flexibility is .84 to .91. The mult-R's for originality are .84, .87, .80 and .73 for activities 4-7 respectively.

The restricted model results parallel those of the full model. In all but one equation, the multiple correlation coefficient dropped by less than one-hundreth of a point. The greatest loss in potential predicability was realized in Activity 7, originality, where a .03 difference was noted. In these restricted analyses, no more than half of the original set of predictors was utilized, with 4 instances of using as few as 5 or 6 predictors. The apparent lack of significant losses in prediction power with a partial set of predictors has important implications for future research. Furthermore, higher cross-validation correlations may be expected because of the reduction in number of predictor variables.

Greater losses in the multiple-R coefficient were detected for the two forced models. Activity 7, flexibility dropped from .84 to .73, and a .13 decline from .73 was noted for the Activity 7, originality forced model. These models were generated, however, because of low cross-validations in their respective full and restricted models, as will be shown soon. This, while lower multiple correlations were obtained, higher cross-validation correlations are expected. One advantage of the particular forced models considered is that they employ only 3 and 4 predictors.

All the multiple correlation coefficients reported are high and significant beyond the .01 level. Before speculating on the true value of these results, however, the validity of the prediction equations will be estimated.

The attenuated cross-validation correlations also appear in Table I. The cross-validation correlations for the first nine equations of the full and restricted models range from .79 to .96. Each is significant beyond the .01 level, but, more importantly, each one indicated that the corresponding equation is capable of excellent prediction. The shrinkage, or difference between the multiple correlation and the cross-validation correlation, is

minimal, never exceeding .10. Thus, the high result level anticipated after considering the multiple regression analyses was in fact achieved for the first nine equations.

Considerable shrinkage was noted in both the full and restricted models for the flexibility and originality dimensions of Activity 7. The attenuated cross-validation correlations of .56 and .48 in the full model and .59 and .62 in the restricted model certainly are at least of moderate value in view of the present state of the art (Page and Paulus, 1968; Archambault, 1969); however, in comparison to the results of equations one through nine, they are somewhat disappointing. Hence, the additional analyses were conducted, and a third model, the forced model, was generated.

As expected, the attenuated cross-validation correlations for the forced models exceeded the corresponding correlations in the first two models. Correlations of .77 and .70 for the flexibility and originality equations in Activity 7 were established. Of course, these results are tenative, and must be tested in a new sample. They represent a goal of minimum stature for future researchers.

The results of the multiple correlation analyses and the corresponding cross-validation correlations must be considered most encouraging. Accurate estimates of the creativity ratings of human judges were achieved by employing computerized content analysis and computer simulation procedures. Perhaps a more significant outcome, however, is the reliability of this automated process. If these same responses were to be rated at a later time, the computer ratings would have a reliability of 1.00. Certainly human judges would not approach this perfection. The success achieved thus far in developing a computerized scoring procedure for creativity tests strongly suggests that similar applications in other areas where open-ended responses are analyzed by human judges are warranted. These other areas include personality and interest tests.

II. Criterion Development

The procedures in this section were employed to establish a reliable criterion for each dimension of each activity. Four judges, rather than a single judge, were used in an effort to maximize the reliabilities of the ratings. The judges were thoroughly trained. In addition, several statistical methods were explored.

The Judges

One professional judge and three educational psychology students were responsible for scoring the responses of the 153 subjects to the TTCT. A professional judge is defined as a judge employed by a test scoring bureau. Of the three trained student judges, one was a first year graduate student, and the other two were completing their third year of undergraduate work.

Procedures for Training Judges

It is assumed that the professional judge was trained in the manner prescribed by Torrance (1966bc). The student judges were trained in the following fashion by Archambault (1969, p. 30):

> To provide uniformity of orientation and to improve inter-scorer reliability, a number of procedures were utilized in the training of the judges.
>
> To give a greater appreciation for the concept of creativity by becoming actively involved in the creative process, each judge was administered the Torrance Tests of Creative Thinking, Verbal Form A. Next, a series of seminars were conducted for the scorers during which the process of creativity and possible problems relating to the scoring procedures were discussed. The scorers were then provided with copies of Torrance's Guiding Creative Talent (1962) and were asked to read selected chapters. Copies of the Torrance Tests of Creative Thinking: Norms-Technical Manual (Torrance, 1966c) and the Torrance Tests of Creative Thinking: Directions Manual and Scoring Guide (Torrance, 1966b) were also provided.
>
> After the literature and manuals had been read, the judges were asked to score a sample set of responses listed in the Scoring Guide. The scorers then met as a group and discussed their rationale for assigning scores to each of the individual responses. Where differences of opinion existed between the judges and the Scoring Guide, the possible reasons for such

differences were analyzed. As a final Activity in the training process, a meeting was arranged between the scorers and Dr. E. Paul Torrance. During this meeting the scorers had the opportunity to raise any unresolved questions emanating from the practice scoring which they had performed.

Additional steps taken to improve reliability included: a) a discussion of the optional amount of time for scoring in any one sitting; b) the provision of a "paste-up" of the scoring manual that enabled the scorers to view one Activity or sub-test at a glance; and c) the scoring of the responses of all subjects to one Activity before proceeding to the next Activity.

Reliability of Judges

Several statistical analyses were performed in an effort to maximize the reliabilities of the creativity ratings. Initially, this writer developed a cycling type of reliability computer program which generated a reliability estimate of the pooled ratings of the judges using analysis of variance techniques (Winer, 1962, pp. 124-132) for all four judges and for all combinations of three judges. The main purpose of this program was to determine if any one judge's ratings should be deleted. The program also generated adjusted reliabilities. These adjusted reliabilities, generally higher than the unadjusted estimates, eliminate the effect of differences in judges' means and should be utilized when the investigator is not willing to accept the assumption of mean homogeneity (Ebel, 1951).

The means and standard deviations for the four judges are presented in Table II. Table III contains the judge inter-correlation matrix. The trained student judges are 1, 2, and 3; judge 4 is the professional scorer.

The results of the cycling program are presented in Table IV. The judge code parameter indicates which judge was not considered in the particular analysis. Judge code "0" indicates that all four judges were considered. Two statements are based on the results. First, a function of all four judges' ratings will constitute the criterion, because generally the highest reliabilities are generated when all four judges were considered.

And second, it is appropriate to utilize the adjusted reliability estimates for the originality dimension.

The judge code "O", adjusted reliabilities for fluency and flexibility are most satisfactory, ranging from .88 to .99 with 6 of 7 above .94. The originality reliabilities, although satisfactory, were somewhat lower. Their range was bounded by .66 and .86. Thus, additional statistical methods were applied to the originality ratings, as suggested by Page and Paulus (1968). This involved factor analysing the raters on the originality scores and determining their factor scores on the first principal component. The factor scores, or some function of them, are then used to differentially weight the raters. As can be seen by examining the results in Table V, the factor scores generated for each scorer were not considerably different. Even if a power function were applied to these loadings, the resulting composite score probably would not differ greatly from a simple average of the scores. Hence, the criterion scores for originality as well as fluency and flexibility were the mean of the four judges' ratings.

III. Generating Parsimonious Prediction Models Through Predictor Stability Analysis

Forced regression models were generated in this study primarily to determine the potential of predicting those creativity scores for which considerable shrinkage was noted in the full and restricted models. A forced model is one in which the researcher selects a partial set of predictors and forces them into the analysis before the remaining variables of the full set are allowed to enter. If the forced model is to differ from the full model, it must also be restricted.

Before considering the process of selecting the forced predictor variables, the rationale for using this type of model will be discussed. In multiple regression analysis, only the full model reflects the present state of the art in whatever field is being studied. The results of restricted and forced

models represent goals to be attained in future research, and each of these models must be applied to a new sample if they are to be recognized as being valid. Thus, when working with models other than the full model, the researcher need not necessarily restrict his efforts to only the development sample. He must realize, however, that the results are tentative and based on the assumptions, however implicit, corresponding to his method of generating the restricted or forced model.

In this study, the forced predictor variables were selected by analysing the correlations between the predictors and the criterion in both the development and cross-validation sample. Only those predictors whose correlations with the criterion did not vary significantly were selected. Referring to this selection technique as predictor stability analysis, this writer recognizes the following aspects: 1) As mentioned earlier, the results obtained must be viewed as representing the future and not the present state of the art. 2) This process does not elminate the inclusion of suppressor predictors. 3) Other researchers have commented on the situation under discussion. Perhaps Page and Paulus best described the problem when they stated (1968, p.53):

> As is well known, however, we should not expect all of this accuracy (high multiple regression coefficients) if we took new essays and applied the discovered beta weightings to them, to predict their human ratings (cross-validation). For any set of scores, or any set of resultant correlations, contains not only true variance associated with the variable, but also a certain amount of error variance, random for the particular subjects concerned, which will not ordinarily be found with a new set of human subjects, or essays. The true variance gives us information which will be subsequently useful. But the error variance is also capitalized upon by the analysis, and a certain portion of the multiple-regression coefficient, and of the contributing beta weights, will spuriously seem to contribute, but will not stand up in a replication.

4) This process, then, is actually an attempt to control the error variance referred to in the above quotation. 5) The stable predictors may be

determined by partitioning only the development sample, but if the regression analysis has already been completed, validation in a new experiment is still necessary. Thus, if a researcher is concerned with the stability of his predictors as related to the criterion, analyses regarding this stability must be performed prior to the generation of the full model. 6) The stability of the correlation may be statistically established by testing for a significant difference between two correlations, as discussed in several texts (e.g., Bruning and Kintz, 1968). However, it should be noted that the test becomes more rigorous by increasing the alpha level. 7) The importance of empirical cross-validation rather than generating a statistical estimate for the behavioral sciences is once again supported. The cross-validation estimates calculated by the Wherry and the Lord-Nicholson formulae are not sensitive to the spurious effect of the error variance and hence are generally optimistic over-estimates.

The attenuated cross-validation results for the forced models in this study were very encouraging. Correlations of .77 and .70 for the flexibility and originality equations in Activity 7 were established. The corresponding full model results were .56 and .48. Of course, the forced model results are tentative, and must be tested in a new sample. They represent a goal of minimum stature for future researchers.

REFERENCES

Archambault, F. A computerized Approach To Scoring Verbal Responses To The Torrence Tests Of Creative Thinking. Unpublished Doctoral Dissertation, The University of Connecticut, 1969.

Ebel, R. Estimation of the Reliability of Ratings. Psychomatrika, 1951, 16, 407-424.

Fisher, G. The SCORTXT Program For The Analysis of Natural Language. Unpublished Manuscript: Bureau of Educational Research, University of Connecticut, 1968.

Guilford, J. P. Creativity. American Psychologist, 1950, 5, 444-454.

Hoepfner, R. Review of Torrence Tests of Creativity. Journal of Educational Measurement, 1967, 4, 191-193.

Lewis, N. (Ed.) Roget's Thesaurus in Dictionary Form. New York: Washington Square Press, 1961.

Mosier, C. The Need and Means of Cross-Validation. Educational and Psychological Measurement, 1951, 11, 5-28.

Page, E. and Paulus, D. The Analysis of Essays by Computer. USOE Final Report, 1968.

Paulus, D. and Renzulli, J. Computer Simulation of Human Creativity Ratings. USOE Proposal for Funds, 1968.

Stone, P. et al. The General Inquirer: A Computer Approach to Content Analysis. Cambridge: The M.I.T. Press, 1966.

Torrance, E. Guiding Creative Talent. Englewood Cliffs, New Jersey: Prentice Hall, 1962.

Torrance, E. Torrance Tests of Creative Thinking. Princeton, New Jersey: Personnel Press, 1966.

Torrance, E. Torrance Tests of Creative Thinking: Directional Manual and Scoring Guide. Princeton, New Jersey: Personnel Press, 1966.

Torrance, E. Torrance Tests of Creative Thinking: Norms and Technical Manual. Princeton, New Jersey: Personnel Press, 1966.

Treffinger, D. and Ripple R. The Effects of Programmed Instruction in Productive Thinking on Verbal Creativity and Problem Solving Among Elementary School Pupils. H.E.W. Final Report, 1968.

Winer, B. Statistical Principles in Experimental Design. New York: McGraw Hill, 1962.

TABLE I

CROSS-VALIDATION RESULTS
INCLUDING MULT-R SUMMARY
ALL MODELS

Activity	Dimension	Full Model			Restricted Model			Forced Model		
		R	r	r Attenuated	R	r	r attenuated	R	r	r attenuated
4	Fluency	.99	.95	.96	.99	.95	.96			
4	Flexibility	.91	.80	.81	.90	.84	.86			
4	Originality	.84	.74	.79	.84	.78	.83			
5	Fluency	.96	.92	.93	.95	.95	.96			
5	Flexibility	.85	.87	.89	.84	.89	.91			
5	Originality	.87	.80	.88	.86	.82	.90			
6	Fluency	.97	.95	.95	.97	.96	.96			
6	Originality	.80	.68	.92	.79	.70	.95			
7	Fluency	.92	.84	.87	.90	.82	.85			
7	Flexibility	.84	.53	.56	.83	.56	.59	.73	.73	.77
7	Originality	.73	.42	.48	.70	.55	.62	.60	.62	.70

**All correlations are significant at .01 level

TABLE II

MEANS AND STANDARD DEVIATIONS FOR JUDGES
TOTAL SAMPLE

Activity	Dimension	Judge 1 $\bar{X}$	Judge 1 SD	Judge 2 $\bar{X}$	Judge 2 SD	Judge 3 $\bar{X}$	Judge 3 SD	Judge 4 $\bar{X}$	Judge 4 SD
4	Fluency	11.80	6.99	11.54	6.68	12.30	7.15	10.68	6.04
4	Flex.	5.84	2.67	5.75	2.72	5.91	2.73	5.28	2.56
4	Orig.	2.37	2.38	8.59	6.31	5.08	4.60	6.99	5.78
5	Fluency	13.90	8.45	16.01	9.82	13.74	8.85	13.97	8.89
5	Flex.	7.39	3.54	7.42	3.55	7.04	3.31	7.15	3.57
5	Orig.	3.84	3.56	11.13	8.45	3.23	3.59	10.72	9.74
6	Fluency	5.33	4.19	5.63	4.19	5.84	4.22	5.64	4.23
6	Orig.	1.32	2.64	4.48	5.94	2.87	4.56	1.33	4.20
7	Fluency	4.75	3.23	4.37	3.24	4.28	2.84	4.39	2.95
7	Flex.	1.99	1.74	2.03	1.99	1.94	1.90	1.86	1.73
7	Orig.	.86	1.07	1.72	1.51	.95	1.36	.91	1.36

TABLE III

JUDGE INTER-CORRELATIONS
TOTAL SAMPLE

Act	Dimension	r for judges x-y					
		1-2	1-3	1-4	2-3	2-4	3-4
4	Fluency	.93	.93	.88	.96	.95	.91
4	Flexibility	.85	.85	.83	.89	.87	.82
4	Originality	.69	.77	.61	.69	.83	.55
5	Fluency	.91	.85	.84	.86	.85	.69
5	Flexibility	.89	.90	.80	.91	.84	.78
5	Originality	.67	.67	.70	.62	.66	.44
6	Fluency	.96	.94	.96	.97	.99	.98
6	Originality	.39	.27	.37	.42	.44	.16*
7	Fluency	.84	.76	.87	.69	.79	.82
7	Flexibility	.58	.78	.66	.59	.58	.74
7	Originality	.47	.50	.27	.53	.51	.27

*Significant at .05 level
All other correlations are significant at the .01 level

TABLE IV

RELIABILITY ESTIMATES FOR ALL JUDGES
AND ALL COMBINATIONS OF THREE JUDGES
USING ANALYSIS OF VARIANCE
TOTAL SAMPLE

Activity	Dimension	Judge Code	Rel.	Adj. Rel.
4	Fluency	0	.98	.98
		1	.97	.98
		2	.96	.97
		3	.97	.97
		4	.98	.98
4	Flexibility	0	.96	.96
		1	.94	.95
		2	.93	.94
		3	.94	.95
		4	.95	.95
4	Originality	0	.76	.86
		1	.83	.87
		2	.64	.77
		3	.65	.83
		4	.61	.81
5	Fluency	0	.95	.95
		1	.92	.92
		2	.92	.92
		3	.94	.95
		4	.95	.95
5	Flexibility	0	.96	.96
		1	.94	.94
		2	.93	.93
		3	.94	.94
		4	.96	.96
5	Originality	0	.64	.80
		1	.58	.75
		2	.43	.67
		3	.66	.79
		4	.45	.74

TABLE IV CONTINUED

Activity	Dimension	Judge Code	Rel.	Adj. Rel.
6	Fluency	0	.99	.99
		1	.99	.99
		2	.99	.99
		3	.99	.99
		4	.98	.99
6	Originality	0	.60	.66
		1	.56	.61
		2	.44	.48
		3	.53	.62
		4	.53	.60
7	Fluency	0	.94	.94
		1	.91	.91
		2	.93	.93
		3	.94	.94
		4	.90	.91
7	Flexibility	0	.88	.88
		1	.84	.84
		2	.89	.89
		3	.82	.82
		4	.85	.84
7	Originality	0	.71	.74
		1	.65	.70
		2	.60	.60
		3	.61	.68
		4	.68	.74

TABLE V

FACTOR ANALYSIS OF JUDGES
ORIGINALITY SCORES
TOTAL SAMPLE

Activity	$\bar{x}$	Eigenvalues	Cumulative %	Loading	r	r Adj.
4	2.37	3.07	77	.87	.76	.86
	8.59	.55	90	.92		
	5.08	.24	96	.86		
	6.99	.15	100	.85		
5	3.84	2.88	72	.90	.64	.80
	11.13	.56	86	.87		
	3.23	.33	95	.80		
	10.72	.22	100	.82		
6	1.32	2.04	51	.71	.60	.66
	4.48	.85	72	.81		
	2.87	.64	88	.62		
	1.33	.47	100	.69		
7	.86	2.28	57	.74	.71	.74
	1.72	.82	78	.85		
	.95	.51	90	.77		
	.91	.39	100	.65		